TCGA Data Coordination Center
4/1/09

Mutation Annotation Format (MAF) File Description

The following data are reported in MAF files:
Somatic mutations

- Missense and nonsense
- Splice site, defined as SNP within 2 bp of the splice junction
- Silent mutations
- Indels that overlap the coding region or splice site of a gene or the targeted region of a genetic element of interest.

SNPs

- Any germline SNP with validation status "unknown" is included.
- SNPs already validated in dbSNP are not included since they are unlikely to be involved in cancer.

Validation
All candidate somatic missense, nonsense, splice site and indels are retested by an independent (orthogonal) genotyping method. If the SNP is confirmed by an independent method, they are deemed valid. Silent mutations may be validated for the purpose of calculating the background mutation rate. No germline (SNP or indel) candidates are processed through validation. However, if the validation process reveals a given candidate somatic variation event to be germline or loss of heterozygosity, those validated data are reported in the validation file.

A *validated somatic mutation* is identified by (Verification_Status=Verified or Validation_Status=Valid) and Mutation_Status=Somatic.

MAF files have a data type of "Mutations". Putative (un-validated) somatic mutations or non-somatic mutations are considered Level 2 data and have controlled access only. Validated somatic mutations (defined above) are considered Level 3 data and open access.

Mutation Annotation Format File Fields

The format of a MAF file is tab-delimited columns. Those columns are described in Table 1. Columns may allow null values (*i.e.* blank cells) and/or have enumerated values.

Table 1 - Mutation annotation format file column headers

Index	MAF Column Header	Description of Values	Null	Fixed
1	Hugo_Symbol	HUGO symbol for the gene, *e.g.* EGFR (HUGO symbols are *always* capitalized)	No	Set
2	Entrez_Gene_Id	Entrez gene ID, *e.g.* 1956	No	Set
3	GSC_Center	Genome sequencing center reporting the variant. One of hgc.bcm.edu, broad.mit.edu, or genome.wustl.edu	No	Yes
4	NCBI_Build	NCBI human genome build number with decimal (e.g. 36.1, 36.2, etc.)	No	Set
5	Chromosome	chromosome number without "chr" prefix, e.g. X, 1, 2	No	Set
6	Start_position	mutation start coordinate (1-based coordinate system)	No	No
7	End_position	mutation end coordinate (inclusive, 1-based coordinate system)	No	No
8	Strand	one of "+" or "-".	No	Yes
9	Variant_Classification	one of Missense_Mutation, Nonsense_Mutation, Silent, Splice_Site_SNP, Frame_Shift_Ins, Frame_Shift_Del, In_Frame_Del, In_Frame_Ins or Splice_Site_Indel	No	Yes
10	Variant_Type	one of SNP, Ins or Del	No	Yes
11	Reference_Allele	the plus strand reference allele at this position	No	No
12	Tumor_Seq_Allele1	tumor sequencing (discovery) allele 1	No	No
13	Tumor_Seq_Allele2	tumor sequencing (discovery) allele 2	No	No
14	dbSNP_RS	dbSNP id (e.g. rs12345) or none or novel	No	Set
15	dbSNP_Val_Status	dbSNP validation status; one of byCluster, bySubmitter, byFrequency, by2hit2allele, byHapmap, none, or unknown	No	Yes
16	Tumor_Sample_Barcode	BCR Aliquot Barcode for tumor sample, i.e. TCGA-SiteID-PatientID-SampleID-PortionID-PlateID-CenterID e.g. TCGA-02-0021-01A-01D-0002-04	No	Set
17	Matched_Norm_Sample_Barcode	BCR Aliquot Barcode for normal sample, e.g. TCGA-02-0021-10A-01D-0002-04 (as opposed to 01A)	No	Set
18	Match_Norm_Seq_Allele1	matched normal sequencing allele or nt (not tested)	No	No
19	Match_Norm_Seq_Allele2	matched normal sequencing allele 2 or nt (not tested)	No	No
20	Tumor_Validation_Allele1	tumor genotyping (validation) allele 1	Yes	No
21	Tumor_Validation_Allele2	tumor genotyping (validation) allele 2	Yes	No
22	Match_Norm_Validation_Allele1	matched normal genotyping (validation) allele 1	Yes	No
23	Match_Norm_Validation_Allele2	matched normal genotyping (validation) allele 2	Yes	No
24	Verification_Status	one of Verified, Wildtype, Unknown	No	Yes
25	Validation_Status	one of Valid, Wildtype, Unknown.	No	Yes
26	Mutation_Status	one of Somatic, Germline, LOH, or Unknown	No	Yes
27	Validation_Method	the assay platform used for the validation call	Yes	No
28	Sequencing_Phase	TCGA Sequencing Phase {1,2,…}	No	Set

Index column indicates the order that the columns are expected. The Null column indicates which MAF columns are allowed to have null values. The Fixed column indicates which MAF columns have specified values: a Fixed value of "No" indicates that there are no specified values for that column; a value of "Yes" indicates that the MAF column requires specific values listed in the Description of Values column; a value of "Set" indicates that the MAF column values come from a specified set of known values (*e.g.* HUGO gene symbols).

Any columns that come after the columns described in Table 1 are optional. Optional columns are not validated by the DCC and can be in any order. The current optional columns are listed in Table 2.

Table 2 - Optional mutation annotation format file column headers

MAF Optional Column Header	Description of Values
Treated_Status	is mutation from a treated sample? {Treated, Non-treated}
Hypermutated_Status	is mutation from a hypermutated sample? {Hypermutated, Non-hypermutated}
COSMIC_COMPARISON(ALL_TRANSCRIPTS)	Comparison of mutation to COSMIC database
OMIM_COMPARISON(ALL_TRANSCRIPTS)	Comparison of mutation to OMIM database
Transcript	Transcript used for annotation
CALLED_CLASSIFICATION	Should be the same as Variant_classification
PROT_STRING	Annotation of mutation effect at the protein level
PROT_STRING_SHORT	Annotation of mutation effect at the protein level (short form. for example, frameshift would be fs)
PFAM_DOMAIN	Annotation of the protein domain that mutation resides in

Additional Information

- TCGA Data Portal http://tcga-data.nci.nih.gov
- TCGA Data Primer: An in depth description of TCGA data enterprise including data classification and organization, how to access the data, and a description of possible ways to aggregate TCGA data. http://tcga-data.nci.nih.gov/docs/TCGA_Data_Primer.pdf
- Data Transfer and Preparation SOP: TCGA standard operating procedures for preparation and transfer of data to the TCGA Data Coordinating Center (DCC). https://gforge.nci.nih.gov/docman/view.php/265/5004/Data_Preparation_and_Transfer_SOP.zip
- NCICB Support http://ncicb.nci.nih.gov/NCICB/support